Unlocking Income Potential: Under or Above 50K

Final Project Report

AI539 - MACHINE LEARNING CHALLENGES

Winter 2024

Oregon State University

Saiyam Maunik Masalia

1. Problem to be solved
    a. Problem Statement:
    The problem presents predicting whether an adult earns over or under $50,000 annually, utilizing various demographic factors such as age, working class, and education. Although the dataset originates from 1994, its relevance extends to comparable datasets. It constitutes a binary classification task aimed at determining whether an individual's income surpasses the $50,000 threshold. Despite its age, the dataset remains applicable for modeling purposes. The task revolves around classifying individuals based on their income level, considering attributes like age, education level, and occupation.

    b. Users/Beneficiaries
    The predictive model for income classification serves a diverse array of users, each benefiting from its insights in distinct ways. Employers and HR departments leverage the model to make informed decisions during recruitment, providing a preliminary assessment of candidates' earning potential based on demographic indicators. Non-profit organizations utilize the model to efficiently allocate resources, ensuring assistance reaches those with incomes below specific thresholds, thereby maximizing impact. Tax authorities and policymakers gain valuable insights into income distribution, facilitating the development of tax laws and policies that promote fairness and economic stability. For job seekers, the model offers a tool to gauge their potential earnings and identify areas for improvement, such as education or work hours, to enhance their income prospects.

2. Dataset Properties
    a. Source – Utilized the data from one of the popular websites for Datasets used for Machine Learning Modelling, UC Irvine Machine Learning Repository. And the dataset was specifically made by a duo, Becker, Barry and Kohavi, Ronny. [1]

    b. Dataset Features (Shape: 48,842 x 15 [including target variable]) [2][3][4]
    i. **Age** (Numerical) - Age of an individual (Range: 17 – 90)

    ii. **workclass** (Categorical) - Employment Status of an individual
        - Unique Values: 'Private', 'Local-gov', '?', 'Self-emp-not-inc', 'Federal-gov', 'State-gov', 'Self-emp-inc',, 'Without-pay', 'Never-worked'

iii. **fnlwgt** (Numerical) - Final Weight. This is the number of people the census believes the entry represents. (Range: 12285 – 1490400) [2]

iv. **education** (Categorical) – The highest Education accomplished.

- Unique values: '11th', 'HS-grad', 'Assoc-acdm', 'Some-college', '10th', 'Prof-school', '7th-8th', 'Bachelors', 'Masters', 'Doctorate', '5th-6th', 'Assoc-voc', '9th', '12th', '1st-4th', 'Preschool'

v. **educational-num** (Numerical) – Same as above but in numerical form (Range: 1 – 16)

vi. **marital-status** (Categorical) - Marital Status of an individual

- Unique Values: 'Never-married', 'Married-civ-spouse', 'Widowed', 'Divorced' 'Separated', 'Married-spouse-absent', 'Married-AF-spouse'

vii. **occupation** (Categorical) - Occupation of an individual

- Unique Values: 'Machine-op-inspct', 'Farming-fishing', 'Protective-serv', '?', 'Other-service', 'Prof-specialty', 'Craft-repair', 'Adm-clerical', 'Exec-managerial', 'Tech-support', 'Sales', 'Priv-house-serv', 'Transport-moving', 'Handlers-cleaners' 'Armed-Forces'

viii. **Relationship** (Categorical) – What relations an individual might have (For e.g. Wife of someone)

- Unique Values: 'Own-child', 'Husband', 'Not-in-family', 'Unmarried', 'Wife', 'Other-relative'

ix. **race** (Categorical) – Race of an individual.

- Unique Values: 'Black', 'White', 'Asian-Pac-Islander', 'Other', 'Amer-Indian-Eskimo'

x. **gender** (Categorical) – Gender of an individual

- Unique Values: 'Male', 'Female'

xi. **capital-gain** (Numerical) - Capital Gains of an individual (Range: 0 – 99999)

xii. **capital-loss** (Numerical) - Capital Loss of an individual (Range: 0 – 4356)

xiii. **hours-per-week** (Numerical) – Number of hours an individual works in a week. (Range: 1 - 99)

xiv. **native-country** (Categorical) – Country of origin for an individual

- Unique Values: 'United-States', '?', 'Peru', 'Guatemala', 'Mexico', 'Dominican-Republic', 'Ireland', 'Germany', 'Philippines', 'Thailand', 'Haiti', 'El-Salvador', 'Puerto-Rico', 'Vietnam', 'South', 'Columbia', 'Japan', 'India', 'Cambodia', 'Poland', 'Laos', 'England', 'Cuba', 'Taiwan', 'Italy', 'Canada', 'Portugal', 'China', 'Nicaragua', 'Honduras', 'Iran',

'Scotland', 'Jamaica', 'Ecuador', 'Yugoslavia', 'Hungary', 'Hong', 'Greece', 'Trinadad&Tobago', 'Outlying-US(Guam-USVI-etc)', 'France', 'Holand-Netherlands'

xv**. income** (Categorical) **- TARGET VARIABLE to be predicted, whether an individual's income lies.**

- Unique Values**:** <=50K, >50K [Binary Classification]

⇨ Class Distribution:

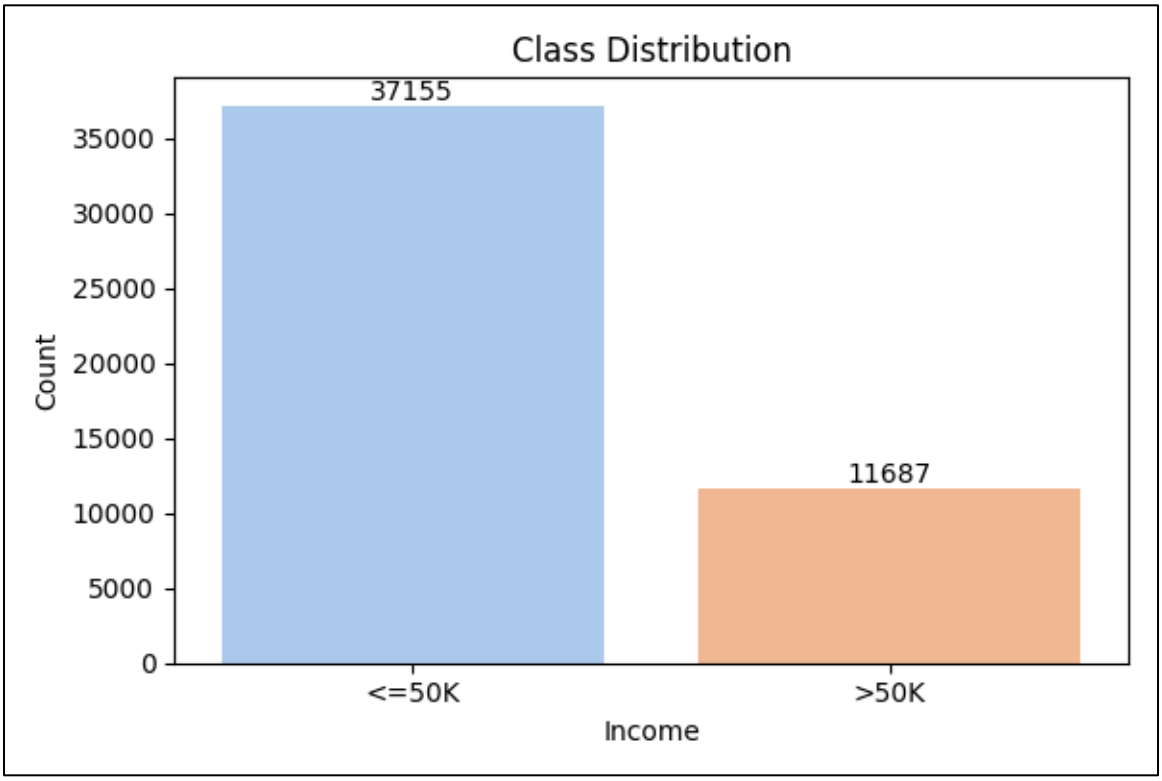

Figure 1: Class Distribution of Target Variable 'income'

⇨ Range of Numerical Columns:

	age	fnlwgt	educational-num	capital-gain	capital-loss	hours-per-week
count	48842	48842	48842	48842	48842	48842
mean	38.6436	189664.135	10.0781	1079.068	87.502	40.422
std	13.711	105604.025	2.571	7452.019	403	12.391
min	17	12285	1	0	0	1
25%	28	117550.5	9	0	0	40
50%	37	178144.5	10	0	0	40
75%	48	237642	12	0	0	45
max	90	1490400	16	99999	4356	99

Table 1: Numerical Columns' Distribution

⇨ Distribution for Numerical Features:

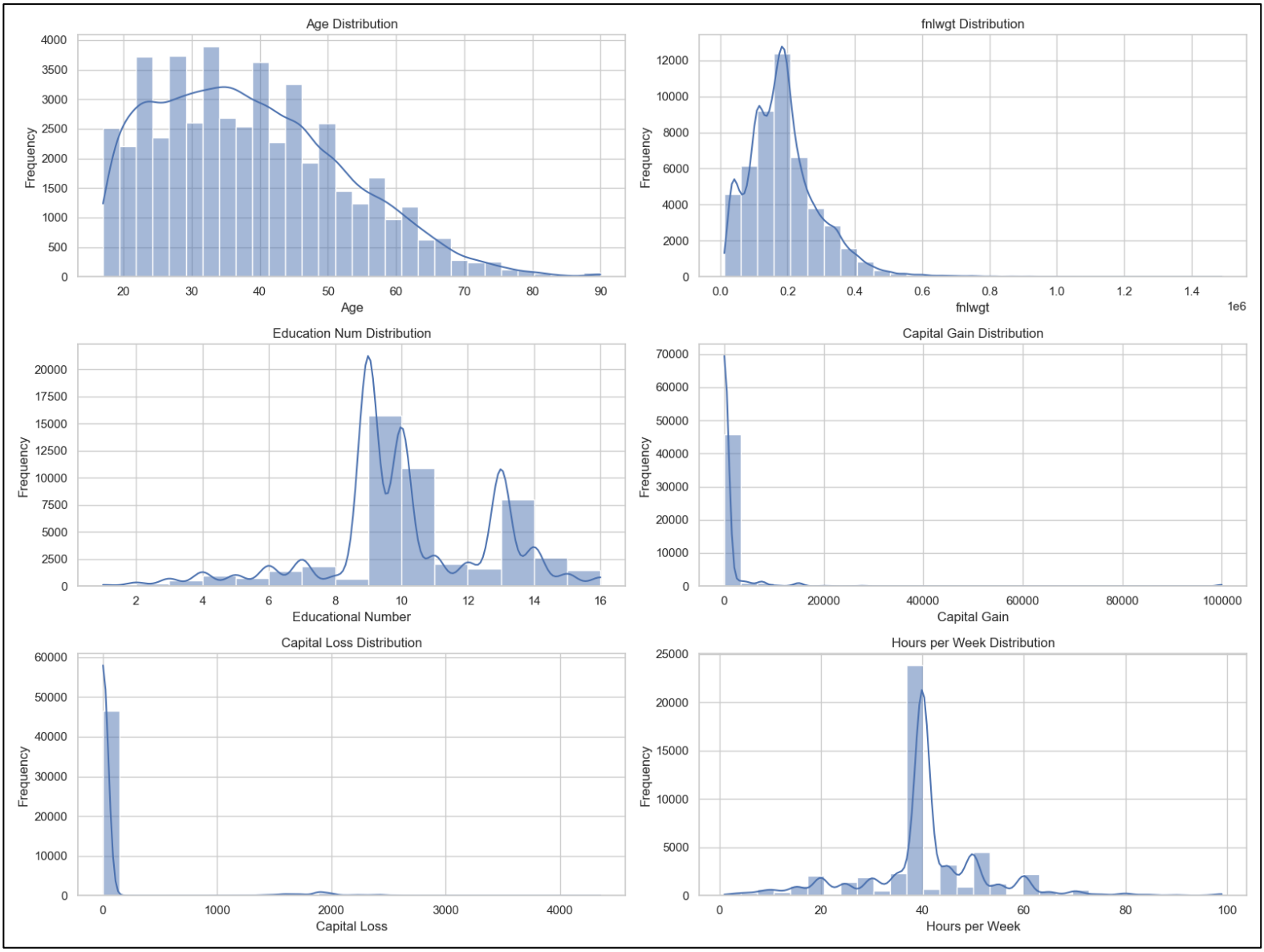

Figure 2: Numerical Features' Distribution

i. Age Distribution: For this Histogram, we can see the majority of individuals are lying in the first half of age groups that is from 20 – 50, it makes sense for people within this age limits to work.

ii. Final Weight Distribution: For this plot, we can see a modal frequency of 12000 for 200000 (0.2 x 10^6) fnlwgt.

iii. Education Number Distribution: The education number distribution appears to be a histogram showing the number of years of education. It has a peak at 8 to 10 years of education, and it tapers off on either side. If we follow, he chronological order of education, we have Preschool, 1st-4th, 5th-6th, 7th-8th, 9th, 10th, 11th, 12th, HS-grad, Some-college, Assoc-voc, Assoc-acdm, Prof-

school, Bachelors, Masters, Doctorate and hence, this suggests that most people in the dataset have a high school diploma or some college education.

iv. Capital Gain Distribution: Here, it appears to be rightly skewed, with most people having small capital gains (majority 0) and a few people having very large capital gains.

v. Capital Loss Distribution: Here, it shows a similar pattern to the capital gain distribution, with most people having small capital losses (again, the majority being 0) and a few people having very large capital losses.

vi. Hours per week Distribution: Here, it appears to be right-skewed, with most people working around 40 hours per week and a few people working many more hours. 40 hours is the standard full-time duration for employees all around the world, so it makes sense to see 40 hours as peak in this graph.

⇨ Distribution of Categorical Features: [From Figure 3]

vii. Workclass Distribution: Majority of individuals are engaged with Private working class which accounts to almost 35000 people. The rest of the people (around 13000) lie in other categories distributed unequally.

viii. Education Distribution: Here, most individuals either completed High School or got a College Diploma. On one side, college diploma is good enough for people to start with some jobs and on either side (for 1996), it was okay for people to start working just on the basis of High School completion. Hence, the results.

ix. Marital Status Distribution: Max two groups, never been married and had a civilian spouse, are in highest frequency.

x. Occupation Distribution: Here, mentioned multiple groups have high frequency, ‘other-service’, ‘prof-specialty’, ‘craft-repair’, ‘Adm-clerical’, ‘Exec-managerial’, ‘Sales’.

xi. Relationship Distribution: Maximum here are people with either no family at all or someone’s husband.

xii. Race Distribution: White dominates all other races in account of more individuals working.

xiii. Sex Distribution: Male: Female = 2:1 ranging in numbers 30000, 15000.

xiv. Native Country Distribution: Main top country is United States.

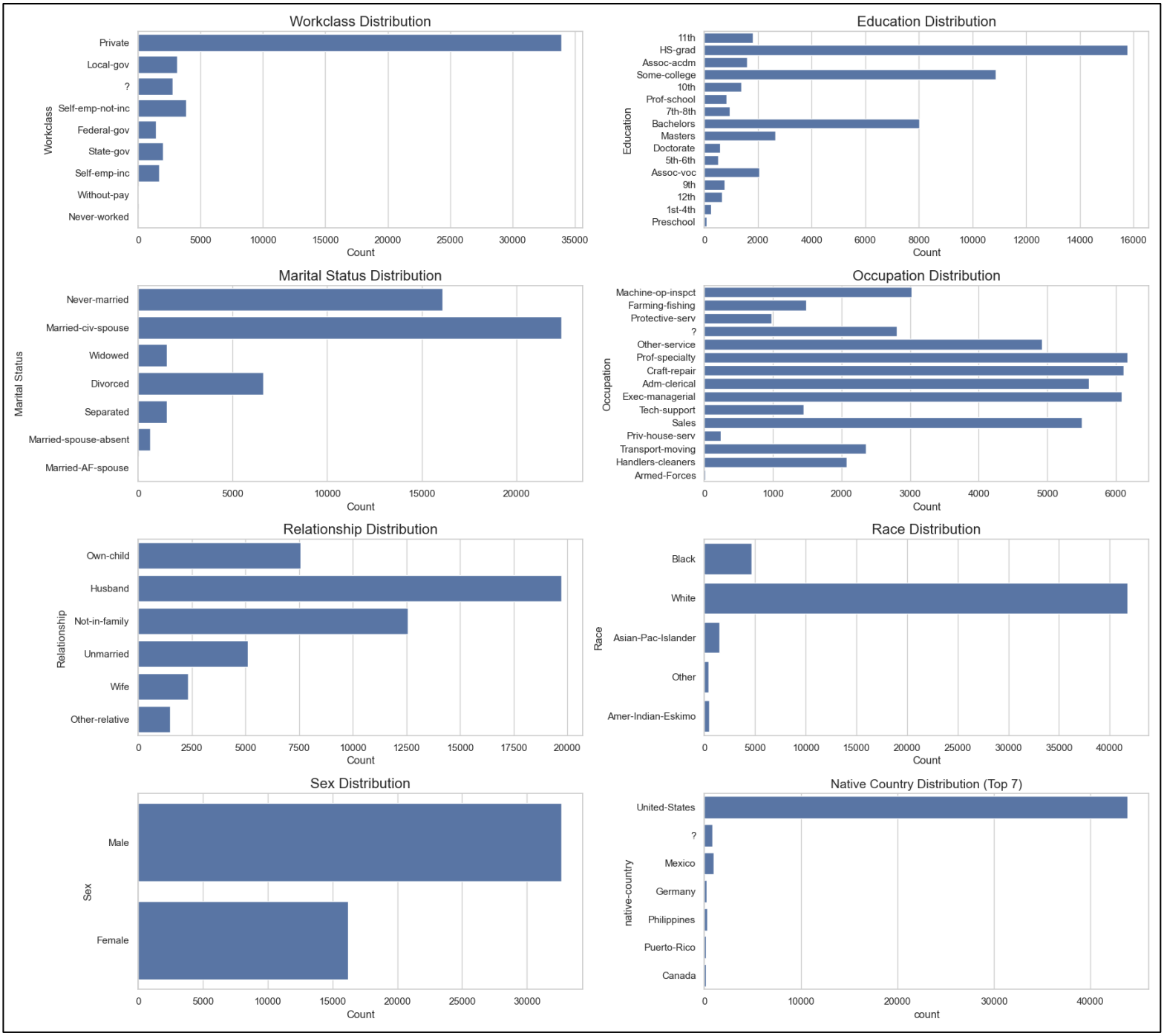

Figure 3: Categorical Feature's Distribution

3. Machine Learning Model:

a. Type of Model Used and Why:

- The model used is XGBoost (Extreme Gradient Boosting). XGBoost is chosen for its effectiveness in handling complex datasets (adult census dataset). XGBoost is a popular choice for structured/tabular data due to its high performance, scalability, and ability to handle missing values and outliers effectively. Additionally, XGBoost typically yields high predictive accuracy and can

handle a large number of features well. For this model, before finalizing XGBoost, RandomForest was used but comparatively XGB gave much better scores, on an average 3% better accurate scores.

b. Strengths and Weaknesses:

Strengths:

- XGBoost generally performs well on tabular data with a large number of features.
- It handles missing values well which is one of the addressed challenges in this project, which is crucial for datasets with real-world noise and incomplete data like the adult census dataset.
- XGBoost is robust to outliers which again is a challenge so to get a boost on metrics, which is evident from the preprocessing steps involving winsorization and imputation.
- For this instance, it was even faster (about 6 secs) than RandomForest which took 8 - 9 seconds.

Weaknesses:

- XGBoost might require more tuning compared to simpler models like logistic regression.
- It could be computationally expensive and memory-intensive for very large datasets.

c. Parameters and Tuning:

- Parameters of the XGBoost model such as `max_depth`, `learning_rate`, `n_estimators`, `subsample`, `colsample_bytree`, etc., should be specified.
- In the provided code, only one parameter `random_state` is explicitly set (random_state = 42). The default values for other parameters are used. Tuning these parameters could potentially improve the model's performance.
- To tune the parameters, techniques like grid search or random search could be employed. Grid search involves exhaustively testing a range of parameter values, while random search randomly samples from a predefined set of values.

4. Evaluation:
   a. Metrics Used: Accuracy (%) and F1 Score were used to measure the performance of the XGBoost classifier. Accuracy provides an overall measure of correct predictions, since we wanted to see how correct our model can be at predicting an individual's income slab. While F1 Score balances precision and recall, especially useful for imbalanced datasets like the one we are using in this case.

   b. Experimental Methodology: The dataset was split into training and testing sets using a ratio of 80:20, ensuring a sufficient amount of data for training while allowing robust evaluation. The XGBoost classifier was trained on the training set and evaluated on the testing set to assess its performance.

   c. Baseline Approach: The XGBoost classifier used on data without any preprocessing other than dropping duplicates served as the baseline approach. It was compared against various strategies (for treating challenges) techniques such as winsorization, mean imputation, binning, dropping rows with missing values, and different resampling techniques like over sampling and SMOTE.

5. Challenges:
   a. Outliers: The dataset exhibits outliers in several demographic indicators, potentially skewing the distribution and affecting the predictive performance of the models. Outliers can distort statistical analyses and model predictions, leading to inaccurate results.

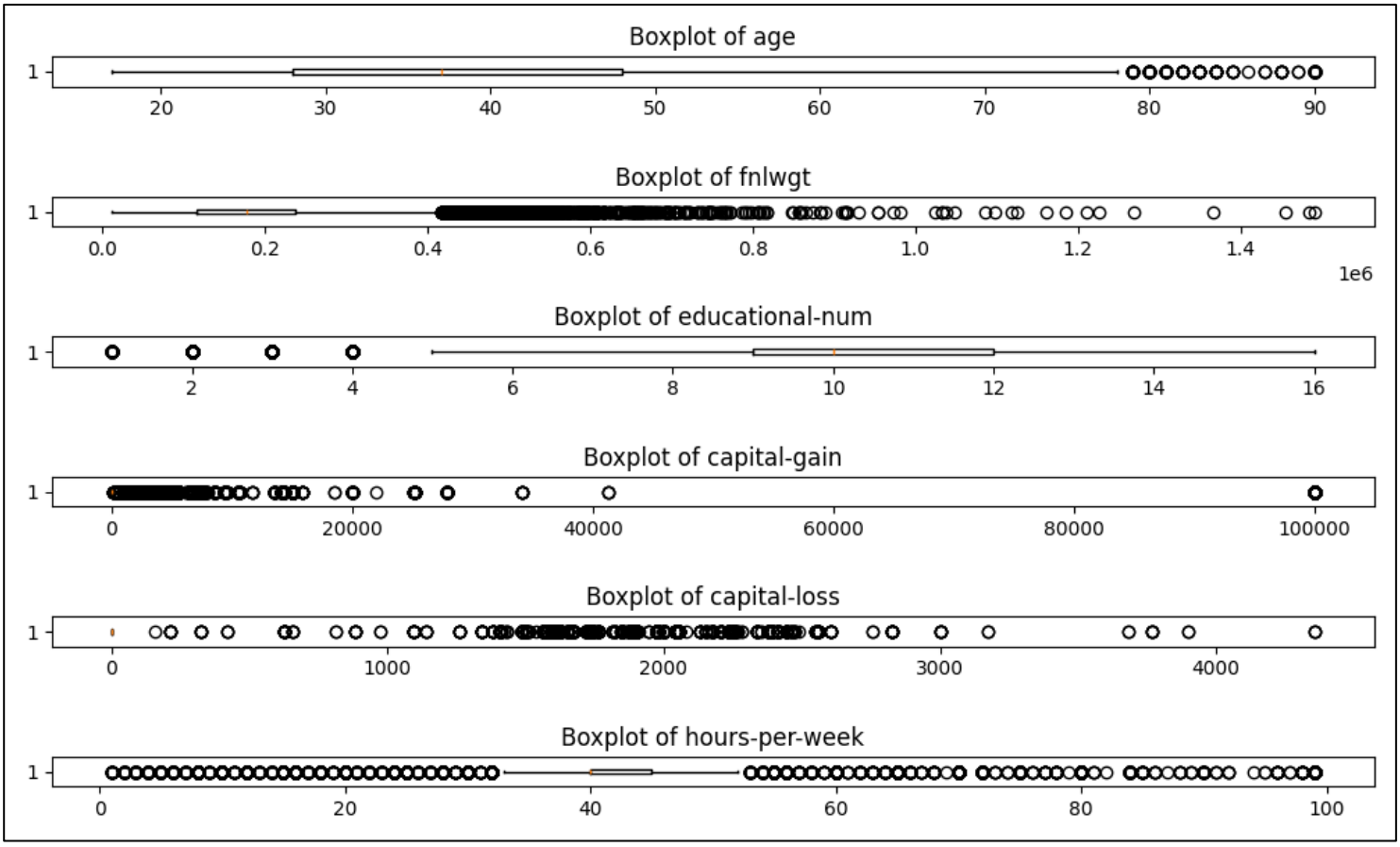

Figure 4: Boxplots on Outliers for Numerical Columns

Strategies :

I. Winsorizing Data: Winsorization involves replacing extreme outliers with values at predefined percentiles (e.g., 5th and 95th percentiles). By capping or flooring extreme values, this technique mitigates the impact of outliers on the analysis while preserving the overall distribution of the data.

II. Imputation (Mean): Imputing missing values caused by outliers with either the mean or median values of the respective features can help maintain data completeness. Mean imputation replaces missing values with the average of the feature, while median imputation replaces them with the median value, both aiming to reduce the influence of outliers.

III. Binning: Binning involves discretizing continuous variables into a finite number of bins or intervals. By grouping similar values together, binning reduces the impact of outliers within each bin and simplifies the data, making it more robust to extreme values.

b. Missing Values: The dataset contains missing values in certain features, which can hinder model training and evaluation, leading to biased results and reduced predictive accuracy. Here we can see an average of 1500 - 2000 missing values in these three features.

Occupation: 2795 missing => MCAR, nothing is causing the data to be missing.
Workclass: 2805 missing => MAR, if occupation is missing, workclass will be too, there's a difference of just 10 rows but that can be considered as MCAR.
Native-Country: 856 missing => MCAR, nothing is causing the data to be missing.
Capital-gain/loss – We can finalize that the 0's we seen are legit.

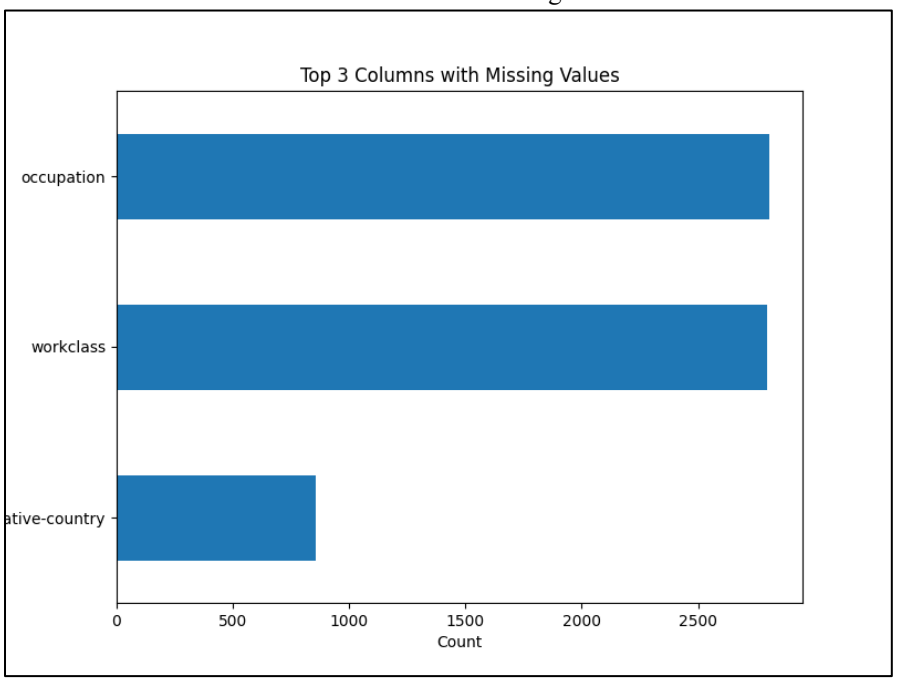

Figure 5: Missing Values

Strategies:

I. Interpolation: Linear interpolation estimates missing values based on surrounding data points, assuming a linear relationship between observed values. By filling in missing values with interpolated estimates, this technique preserves the overall trend of the data while completing the dataset.
II. Mean Imputation: Imputing missing values with statistical measures such as the mean of the respective features provides a simple yet effective way to handle missing data. Mean imputation replaces missing values with the average of the feature, median imputation replaces them with the median value, and mode imputation replaces them with the most frequent value, ensuring data completeness.
III. Dropping Data Rows: Removing rows with missing values from the dataset ensures that the model learns from complete data. While this approach may lead to information loss, particularly if the missing data is not completely at random, it helps prevent bias and ensures the integrity of the analysis.

c. Class Imbalance: The target variable exhibits class imbalance, with one class significantly outnumbering the other, posing a challenge for model training and evaluation. It's clearly visible from Figure 1 where there is a 3:1 Class imbalance where <=50K is dominant. (Roughly 37K to 12K ratio) [Figure 1]

Strategies:

I. Over Sampling: The class imbalance in the dataset is addressed by applying the RandomOverSampler technique, which duplicates instances from the minority class to balance the distribution. This ensures that the model receives more samples of the minority class during training, improving its ability to make accurate predictions for both classes.

II. SMOTE (Synthetic Minority Over-sampling Technique): To overcome the class imbalance, SMOTE generates synthetic samples for the minority class based on feature similarity with existing instances. By creating artificial data points, SMOTE effectively expands the minority class representation, enabling the model to learn more robust decision boundaries and improve its generalization capability.

III. Class Weights: Adjusting class weights in the Random Forest model assigns higher importance to minority class instances, prioritizing their correct classification and penalizing misclassifications more heavily. So, again, while training we can maintain a balance between the two classes.

6. Results:

Preprocessing	Accuracy (%)	F1 Score
Base	87.026	0.704
Winsorization 5/95	84.864	0.656
Mean Imputation in Outliers	84.136	0.629
Binning in Outliers	85.776	0.67
Dropping NA	86.563	0.704
Mean Imputation in NA	87.344	0.709
Linear Interpolation	84.341	0.595
RANDOM Over Sampling	83.87	0.713
SMOTE	85.95	0.702
Class Weights	85.263	0.657

Table 2: Results

a. Base: Original Data after removing duplicate rows was used for fitting to the model, no other pre-processing was done other than encoding. And yet, surprisingly, gave second best score, 87.026%.

b. Challenge 1: Outlier
In our experiment, neither Winsorization nor Binning, 84.864% and 85.776% respectively, outperformed the base model (87.026%), indicating limited effectiveness in mitigating outlier influence. Winsorization replaces extreme values (Beyond limits of 5/95 percentile

mark as stated) with less extreme ones (Values lying on 5/95 percentile mark), while Binning categorizes data into intervals (bins of size 10), yet both strategies fell short in improving model performance. Additionally, the lowest one for Outliers, Mean Imputation, imputes outliers by means of rest of the values, also showed decreased performance. This underscores the complexity of outlier handling and suggests that further investigation. Collectively, treating outliers using and of the three strategies didn't help the model.

c. Challenge 2: Missing Values
Mean imputation proved to be the most effective strategy for handling missing values, yielding the highest accuracy (87.344%) and second best F1 score (0.709) in our experiment. This approach fills missing values with the mean of the feature, preserving the overall distribution of the data. Next, Dropping NA, although straightforward, resulted in slightly lower performance compared to Mean Imputation, only possible reason is the loss of valuable information. Lastly, Linear Interpolation, while useful, exhibited lower performance, indicating that interpolating missing values linearly may not adequately capture the underlying patterns in the data.

d. Challenge 3: Class Imbalance
In addressing imbalanced data, various techniques were explored, including Random Oversampling, SMOTE, and Class Weights. Interestingly, Random Oversampling yielded the highest F1 score (0.713), indicating its effectiveness in handling imbalanced classes by replicating minority class samples. SMOTE (85.95%) slightly outperformed Random Oversampling (83.87%) in terms of accuracy, showcasing its capability to generate synthetic samples more effectively. Class Weights also exhibited promising results, indicating that assigning different weights to classes during model training can effectively balance the impact of minority and majority classes on model performance. These findings highlight the importance of tailored approaches to rebalance imbalanced datasets for optimal classification performance.

7. Reflection
   A. What surprised you about your results?
   Ans – I did not expect Base to still top off the accuracies of all 10 strategies (after mean imputation in NA). I tried some preprocessing like reducing unique values for each column and yet, scores would vary but consistently "Base" would give best scores. Same case for RandomForest too which was used earlier. And, the most shocking part, dropping columns "capital-gain" & "capital-loss" showed dropped in accuracies. These are the two columns with highest number of zeros, which as pointed in presentation didn't give any idea on whether that was the incurred gain/loss or whether it was lost information.

   B. Show your model's performance results to someone else in the class and ask them to identify one aspect of the results that seems surprising, or they would like to understand better. Describe their feedback.
   Ans – "*In my opinion, it's actually quite surprising to see that Mean Imputation in NA has the highest Accuracy and F1 Score. I mean if you were to think of it, it is actually ironic, especially since more sophisticated methods like SMOTE and over sampling, which are specifically designed to address class imbalance, did not lead to the highest scores and in fact falls short of imputation by mean in NA.*

I mean, in a normal case scenario, one would expect techniques specifically designed to handle class imbalance, like SMOTE or over sampling, to improve the model's F1 Score considering how F1 Score is a metric that takes into consideration both precision and recall and is ideally supposed to be the least biased on imbalanced datasets. Which is why I am quite intrigued, how these techniques did not outperform the simpler mean imputation in terms of F1 Score.

But again, c'est la vie…." – Jibin Yesudas Varghese.

C. What did you have to (or choose to) change (from your original plans)?
Ans – Originally, I was experimenting with Logistic Regression and Random Forest and in which I was going to finalize the latter. But after researching a bit, I thought of adding XGBoostand comparing with existing algorithms I used before (SVM, KNN, and so on). And luckily, it gave better scores. As mentioned, on average 3% better accuracies than RF did, one of the possible reasons would be that XGB is better at mitigating overfitting, and this is moderately complex dataset.

D. If you had more time, what additional investigation would you do with this data set? (What are your unanswered questions?)
Ans - Given more time, I'd prioritize conducting an extensive hyperparameter search for XGBoost, RandomForest, and other models to maximize performance while exploring interactions with preprocessing techniques. Additionally, I'd assess the robustness of preprocessing methods across demographic subsets /features to ensure generalizability and for this, I would like to reduce the number of unique values, maybe not drastically but whichever yields good results. Further, I'd deepen the analysis of feature importance to uncover key predictors of income, refining feature engineering. This is a converse for features 'capital-gain', 'capital-loss' which had majority of values as zeros and yet it had some significance. And finally, why is it that still after all pre-processing and changes I got 'Base' as one of the best accuracy values.

8. Extra Credit
User Feedback:
"You've navigated the preprocessing challenges quite well. Dealing with outliers and missing data is not an easy task, and your exploration of techniques like Winsorization and mean imputation shows good selection. Maybe you can try refining your approach, especially in addressing class imbalances. However, there are still some areas where additional attention could yield even better results. Addressing class imbalances, for instance, might require a bit more experimentation to find the perfect balance. And don't forget to explore deeper into feature engineering. Additionally, ensure that you're using a fresh and updated dataset". – Mihir Patel, a roommate of mine (Sorry, I couldn't find a potential or current user, I asked him since we share the interest in ML, and he could give some helpful advice)

References:

1. Becker,Barry and Kohavi,Ronny. (1996). Adult. UCI Machine Learning Repository. https://doi.org/10.24432/C5XW20.
    a. URL - https://archive.ics.uci.edu/dataset/2/adult from *UC Irvine Machine Learning Repository*
    b. Refer this URL - https://oregonstate.app.box.com/folder/254409913554?s=nqu666k2d9me2tjrz34dn4g22h6agi3y

2. https://www.kaggle.com/code/yashhvyass/adult-census-income-logistic-reg-explained-86-2

3. https://cseweb.ucsd.edu/classes/sp15/cse190-c/reports/sp15/048.pdf

4. https://www.kaggle.com/code/jieyima/income-classification-model